Do observations have any role in science policy studies? A reply

van den Besselaar, P.A.A.; Heyman, Ulf; Sandstrom, Ulf

published in
Journal of Informetrics
2017

DOI (link to publisher)
10.1016/j.joi.2017.05.022

document version
Publisher's PDF, also known as Version of record

citation for published version (APA)
van den Besselaar, P. A. A., Heyman, U., & Sandstrom, U. (2017). Do observations have any role in science policy studies? A reply. *Journal of Informetrics*, *11*(3), 941-944. https://doi.org/10.1016/j.joi.2017.05.022

Journal of Informetrics 11 (2017) 941–944

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Correspondence

Do observations have any role in science policy studies? A reply

1. Introduction

In Van den Besselaar et al. (2017) we tested the claim of Linda Butler (2003) that funding systems based on output counts have a negative effect on impact as well as quality. Using new data and improved indicators, we indeed reject the claim of Butler. The impact of Australian research improved after the introduction of such a system, and did not decline as Butler states. In their comments on our findings, Linda Butler, Jochen Gläser, Kaare Aagaard & Jesper Schneider, Ben Martin, and Diana Hicks put forward a lot of arguments, but do not dispute our basic finding: citation impact of *Australian research went up*, immediately after the output based performance system was introduced.

The response by our commenters boils down to the following points:

1. Butler never claimed that quality went down (only impact).
2. Butler's study is better than ours as she has (i) more background knowledge and (ii) more accompanying evidence.
3. The timing of the policy intervention and its effects.
4. Results of studies are unimportant in the social sciences, as we never can end a discussion using empirical evidence.
5. Causality cannot be claimed by us on the positive effects.
6. The model should be much more complex.

We will address these issues below. But before doing so, we start with two general remarks. First, our paper is merely critical and shows that the Butler claims are wrong, and that policy lessons based on those findings are unfounded. This is important, as Butler's papers are often used to criticize evaluation and funding systems (Hicks et al., 2015; Wilsdon et al., 2015) and incentives (Stephan, 2012). Secondly, based on that, we argue that stimulating productivity is important – as we have indicated elsewhere at the individual level (Sandström and van den Besselaar, 2016). Successfully implementing such an output stimulating policy may not be trivial (Sanz Menendez et al., 2008), and depends on more characteristics of the funding and evaluation ecology (Sandström et al., 2014; Van den Besselaar & Sandström, in preparation). However, that discussion falls outside the scope of our paper.

2. Impact versus quality

Butler (in this issue) states: "One thing I have never done, and will never do, is equate low impact (whether determined by citation counts for individual publications or for journals) with lack of quality. I am extremely careful to discuss "*impact*", not "*quality*", and the latter term only appears in my publications in relation to policy statements." But Butler may have forgotten her original interpretation of the findings on Australia (co-authored with Gläser): ". . . the introduction of publication counts into funding formulas was followed by the emergence of a *gap between quantity and quality of publications*" (Gläser et al., 2002).

It is obvious that Butler and Gläser use impact here as a proxy for quality. But interestingly, when they link impact with quality it is acceptable, but when we do the same it is wrong (as it does not fit our opponents' opinions). Although we understand that there are more quality dimensions than citation impact – for scholarly quality at the level of national science systems, impact can safely be used. In fact, the impact indicator we use, the share of top 10% highly cited papers, does this much better than the average citation impact used by Butler (Waltman & Van Eck 2015).

One more issue related to the impact indicators should be added here. Butler puts a lot of emphasis on the impact factor of the journals the Australian researchers started to publish in. A main issue in her argument was that the share of lower impact journals increased, which she saw as a sign of the decreasing quality of Australian research. We argued, referring to Seglen (1994, 1997) that the impact factor in fact is not useful for studying the impact and quality of research. A recent paper extending the small case study of Seglen to a much larger set of researchers confirms our point (Zhang et al., 2017).

http://dx.doi.org/10.1016/j.joi.2017.05.022

3. The role of background knowledge

Of course, background knowledge and accompanying evidence may be relevant if used productively. However, much background and insider knowledge may also work in the opposite direction, and lead to biased observations and conclusions. Accompanying evidence would be great if the main conclusions are correct. Unfortunately, they are not in this case and the accompanying evidence now may mislead readers more than that it gives adequate directions.

This problem emerges when too much engagement with the case may hinder a neutral view. Looking back, one may see that Butler's results were expected by many and were seen as a corroboration of widely shared expectations. These expectations had been explicitly formulated in reports issued by the Science Policy Research Unit (e.g., Geuna & Martin 2001). To some extent, one may say that the field was waiting for these results to be presented, and when they were presented not many looked into the methods and the data. At that time no one took any attempt to replicate the results. They were very quickly taken for granted and carved in stone. When Ben Martin in this issue concludes that the study of Linda Butler is more convincing than ours, this is more than expected: in contrast to our findings, Butler's results do fit in his expectations – even if they are empirically wrong. That also holds for Aagaard and Schneider (this issue; see also Schneider et al., 2016), who also do not see that Butler's conclusions are wrong, although they present the evidence for that (Van den Besselaar and Sandström, in press). Our commenters seem to prefer to keep the wrong belief alive.[1]

4. The timing of the policy intervention

In this context it may be useful to say a few words about the *timing* of the policy intervention in Australia, and the moment one would expect effects. Especially Linda Butler and Diana Hicks (both in this issue) put a lot of emphasis on this, but also the other commenters mention this issue. Butler and Hicks argue that our paper is completely missing the point as we identify these moments wrongly. Let's assume that we did. Would this have impact on our conclusions? As Figures 3 and 4 in our paper clearly show, the increase of impact started already in 1992, which is immediately after 1991, the year of the introduction of the new policy according to Butler. The issue of "what year" has no effect on our findings of increasing impact.

5. Can observations be decisive?

Ben Martin and Jochen Gläser (both in this issue) seem to be somewhat postmodern when they claim that social scientists will always find contradicting results, and seem to imply that research cannot lead to solving the contradiction. However, the examples used to support their claim are in fact illustrations of the opposite. (i) The market pull-technology push discussion has been resolved, as researchers tried to find out why these on first sight contradicting findings could occur. (ii) And in case of the current discussion: Butler simply did not show that a declining impact followed on the policy interventions. So it is not about contradicting findings at all. Interestingly, none of our commenters explicitly defends the claim of Butler that impact went down. This is wise, but instead of drawing conclusions from that, we observe a move to a meta level debate. More generally, we dislike a move to 'epistemological arguments' when one disagrees at the empirical level; this is non-productive and hinders advancing knowledge about the phenomena under discussion.

6. The causality issue

The causality argument suffers from a similar problem as the quality versus impact argument: Gläser and Butler (Gläser et al., 2002) stated that "A recent investigation of Australia's scientific output gives raise to concerns about the continued use of formulas in their existing forms. It documents a significant increase in the country's journal output, accompanied by a worrying decrease in the relative international impact of these publications as measured by citations (Figures 1 and 2). *The timing of this productivity increase in relation to the introduction of funding formulas suggests that there is a* causal *relationship*." (italics by us).

When we make the argument that the timing of the increase of impact suggests a causal relation, it is not acceptable any more to Gläser (this issue), clearly using different standards. Aagaard & Schneider (this issue) also warn that we too easily make causal interpretations. But in their own work they stated the following: "Consequently, the answer to the research question: 'What happens at the aggregated national output and impact levels when institutional funding is linked to differentiated publication counts?' is—for the Norwegian case: (1) publication activity goes up, (2) impact remains stable, and (3) there is no indication of a shift of journal publication activities towards the lowest-impact journals in the WoS" (Schnieder et al., 2016).

One may claim that this wording is not a causal statement, but then the whole argument becomes meaningless. And what is meant by: "What happens when . . ."? We would say: causality!

[1] The role of opinions in research is always a problem and the research system has developed a number of institutions to handle those. Various forms of replication studies are one good option. Another would be to have peer review only of the method and the data but not the substantial results (the model used by *PLOS ONE*), as some argue that this is where it goes wrong (*Research Professional* September 9, 2016).

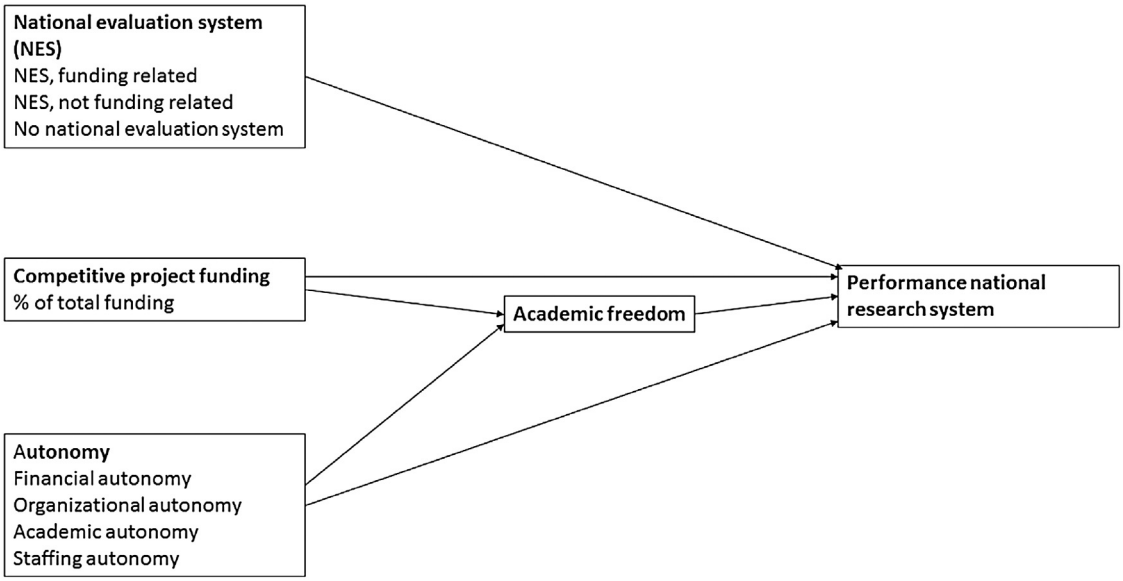

Fig. 1. Explaining the efficiency of national science systems (Van den Besselaar & Sandström, in preparation).

7. The need for more complex models

This brings us to the last point: Gläser (in this issue) repeats that a model explaining research performance needs to be more complex. We agree, but that was not the aim of our paper in this issue, which was to replicate Butler's (2003) study and to show that her conclusions are wrong.

But in an older paper, we developed a more complex model of the relation between funding, evaluation and performance, and introduced a concept of 'funding ecologies'. We used this for comparing impact of different funders (Van den Besselaar and Sandström, 2013). In a follow-up study, we further developed a model to relate performance of the research system to funding ecologies. We explained differences in performance, defined as the ratio between the increase in top cited papers and the increase of total research funding, with several independent variables reflecting funding and evaluation systems (Fig. 1):

(i) The existence of national research evaluation systems (three categories: no national evaluation system; a national evaluation system without funding implications; a system with funding implications).
(ii) The share of competitive project funding in total research funding.
(iii) The level of university autonomy in four dimensions.
(iv) The level of academic freedom, which we assumed to be influenced (negatively) by the level of university autonomy and (positively) by the availability of competitive funding.

We did find evidence that different performance based evaluation systems have different effects on performance, and that other (institutional) variables play a role too (Sandström et al., 2014; Van den Besselaar & Sandström, in preparation).

8. Conclusions

It is important to test the findings of Butler about Australia – as these findings are part of the accepted knowledge in the field, heavily cited, often used in policy reports, but hardly confirmed in other studies. We found that the conclusions of Butler are wrong, and that many of the policy implications based on it simply are unfounded. In our study, we used better indicators, and a similar causality concept as our opponents. And our findings are independent of the exact timing of the policy intervention.

Furthermore, our commenters have not addressed our main conclusions at all, and some even claim that observations do not really matter in the social sciences. We find this position problematic – why would the taxpayer fund science policy studies, if it is merely about opinions? Let's take science seriously – including our own field.

References

Butler, L. (2003). Explaining Australia's increased share of ISI publications—the effects of a funding formula based on publication counts. *Research Policy, 32*, 143–155.
Geuna, A., & Martin, B. (2001). University research evaluation and funding: An international comparison. In *SPRU electronic working paper series No. 71*.

Gläser, J., Laudel, G., Hinze, S., & Butler, L. (2002). Impact of evaluation-based funding on the production of scientific knowledge: What to worry about and how to find out. In *Report: Expertise for the german ministry for education and research*.
Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden manifesto for research metrics. *Nature*, *22*(April).
Research Professional (2016), Trial set up to test publisher's action against publication bias (September 9, 2016).
Sandström, U., & van den Besselaar, P. (2016). Quantity and/or quality? The importance of publishing many papers. *PLoS One*, *11*, e0166149. http://dx.doi.org/10.1371/journal.pone.0166149
Sandström, U., Heyman, U., & van den Besselaar, P. (2014). The complex relationship between competitive funding and performance. In *Proc. science & technology indicators conference*.
Schneider, J. W., Aagaard, K., & Bloch, C. W. (2016). What happens when national research funding is linked to differentiated publication counts? A comparison of the Australian and Norwegian publication-based funding models. *Research Evaluation*, *25*(3), 244–256.
Seglen, P. O. (1994). Causal relationship between article citedness and journal impact. *Journal of the American Society for Information Science*, *45*, 1–11. http://dx.doi.org/10.1002/(SICI)1097-4571(199401)45:1<1:AID-ASI1>3.0.CO;2-Y
Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ*, *314* http://dx.doi.org/10.1136/bmj.314.7079.497
Stephan, P. (2012). Perverse incentives. *Nature*, *484*(April (212)), 29–231.
Van den Besselaar, P., Sandström, U., et al. (2013). The effects of funding modes on the quality of knowledge production. In Juan Gorraiz, & Edgar Schiebel (Eds.), *Proc ISSI 2013* (pp. 664–676).
Van den Besselaar P., Sandström U., Counterintuitive effects of incentives? *Research Evaluation* (In press).
Van den Besselaar P, & Sandström U. *Funding modes, evaluation systems and the performance of national research systems*. (in preparation).
Van den Besselaar, P., Heyman, U., & Sandström, U. (2017). Perverse effects of output-based research funding? Butler's Australian case revisited. *Journal of Informetrics*, *11*, 905–918.
Waltman, L., & Van Eck, N. J. (2015). Field-normalized citation impact indicators and the choice of an appropriate counting method. *Journal of Informetrics*, *9*(4), 872–894.
Wilsdon, J., et al. (2015). *The metric tide: Report of the independent review of the role of metrics in research assessment and management*. http://dx.doi.org/10.13140/RG.2.1.4929.1363
Zhang, L., Rousseau, R., & Sivertsen, G. (2017). Science deserves to be judged by its contents, not by its wrapping: Revisiting Seglen's work on journal impact and research evaluation. *PLoS One*, *12*(3), e0174205. http://dx.doi.org/10.1371/journal.pone.0174205

Peter van den Besselaar*
Department of Organization Sciences & Network Institute, Vrije Universiteit Amsterdam, Netherlands

Ulf Heyman
Uppsala University, Uppsala, Sweden

Ulf Sandström
Dept. INDEK, KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden

* Corresponding author.
E-mail addresses: p.a.a.vanden.besselaar@vu.nl (P. van den Besselaar), ulf.heyman@uadm.uu.se (U. Heyman), ulf.sandstrom@indek.kth.se (U. Sandström)
Available online 9 July 2017